Comparing Direct and Indirect Temporal-Difference Methods for Estimating the Variance of the Return

Craig Sherstan, Dylan R. Ashley*, Brendan Bennett*, Kenny Young, Adam White, Martha White, Richard S. Sutton

Reinforcement Learning and Artificial Intelligence Laboratory, University of Alberta

Background

In a Markov Decision Process the return is defined to be the discounted sum of future rewards:

$$G_t = R_{t+1} + \gamma_{t+1} R_{t+2} + \gamma_{t+1}\gamma_{t+2} R_{t+3} + \dots$$

The λ-return provides a bias-variance trade-off controlled by a new hyperparameter λ and often is easier to learn while being just as useful (we use J to denote the expected value of the return from a state, or the *value* of a state):

$$G_t^\lambda = R_{t+1} + \gamma_{t+1}(1-\lambda_{t+1})\, J_t(S_{t+1}) + \gamma_{t+1}\lambda_{t+1} G_{t+1}^\lambda$$

Note that the λ-return is exactly equal to the return when λ is kept at 1.

Motivation

We might want to learn the variance of the return as

- it lets us take risk into account when making decisions [3],
- we can use it for hyperparameter tuning [5],
- we can use it to guide exploration [4], and
- we can use it to improve our representation [1].

Direct Method

We present a new, direct approach [2] that works by learning the expected value and variance of the return using TD(λ). The direct method uses the following variance estimator:

$$\begin{aligned}
\bar{\gamma}_{t+1} &\leftarrow \gamma_{t+1}^2 \\
\bar{R}_{t+1} &\leftarrow \delta_t^2 \\
\bar{\delta}_t &\leftarrow \bar{R}_{t+1} + \bar{\gamma}_{t+1} J_t(s') - J_t(s) \\
V_{t+1}(s) &\leftarrow V_t(s) + \bar{\alpha}\,\bar{\delta}_t
\end{aligned}$$

Both the direct and the existing indirect method [5] can be viewed as a network of learners:

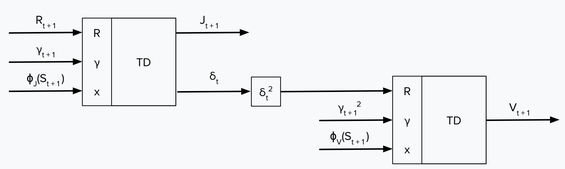


Figure 1: The direct method

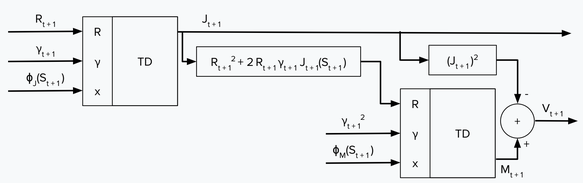


Figure 2: The indirect method

Intuition

For a random variable X, the variance of X can be expressed as $\mathbb{E}[(X - \mathbb{E}[X])^2]$. Similarly, the variance of the λ-return can be expressed in terms of temporal difference (TD) errors:

$$\left(G_t^\lambda - J(S_t)\right)^2 = \left(\sum_{n=0}^{\infty} \delta_{t+n} \prod_{k=1}^{n} \gamma_{t+k}\lambda_{t+k}\right)^2 \approx \sum_{n=0}^{\infty} \delta_{t+n}^2 \prod_{k=1}^{n} \gamma_{t+k}^2 \lambda_{t+k}^2$$

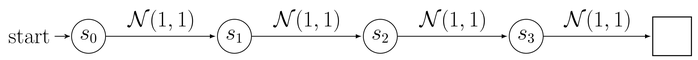
This has the form of a Bellman equation:

$$\mathrm{Var}[G_t^\lambda] \approx V(S_t) = \mathbb{E}[\delta_t^2 + \gamma_{t+1}^2 \lambda_{t+1}^2 V(S_{t+1})]$$

and so can be approximated using temporal difference methods.

Tabular Results

We compare both methods on a chain domain first:



For this domain we highlight the resilience our method to differences in step sizes between the value and variance learners:

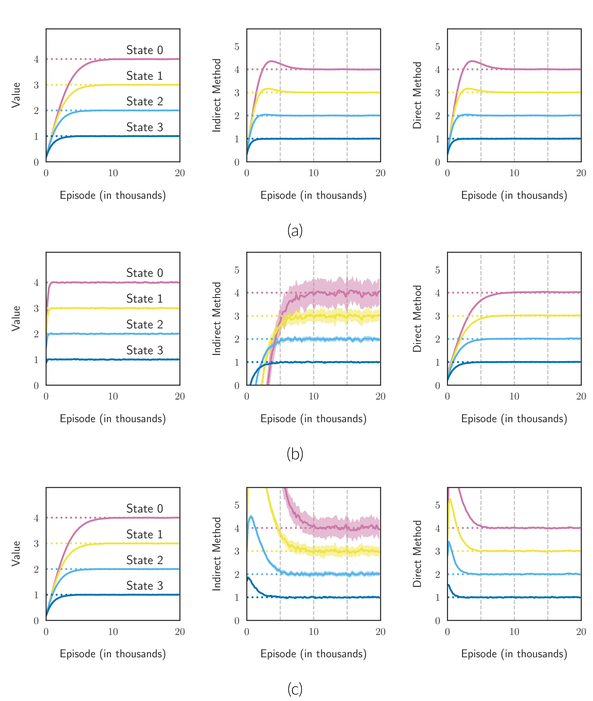
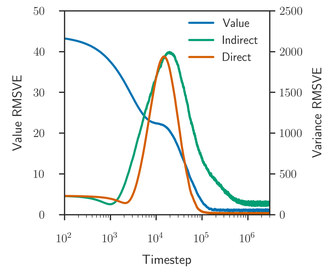


Figure 3: Learning on chain with **a)** step-sizes equal ($\alpha = \bar{\alpha} = 0.001$), **b)** value step-size larger, ($\alpha = 0.01, \bar{\alpha} = 0.001$), **c)** variance step-size larger, ($\alpha = 0.001, \bar{\alpha} = 0.01$)

Linear Function Approximation Results

We experiment on a function approximation scheme overlaying a tabular random walk:

For state S_i the value estimator uses $\phi_t = [1, i]$ and the variance estimator uses $\phi_t = [1, i, i^2]$. Here we highlight the faster learning and better convergence of our method:



Conclusions

In general, the direct method estimates the variance of the return at least as well as the indirect method and typically better. In particular the direct method

- is more stable, even during the transient period before the value function has converged,
- converges faster and is more robust to errors in the value estimator,
- can learn robustly over a wider range of hyperparamters (for example, different step-sizes for the value and variance estimators, or longer eligibility traces), and
- exhibits substantially better performance under linear function approximation.

References

[1] Mark B Ring.
Child: A first step towards continual learning.
Machine Learning, 28(1):77--104, 1997.

[2] Craig Sherstan, Dylan R Ashley, Brendan Bennett, Kenny Young, Adam White, Martha White, and Richard S Sutton.
Comparing direct and indirect temporal-difference methods for estimating the variance of the return.
In *34th Conference on Uncertainty in Artificial Intelligence*, 2018.

[3] Aviv Tamar, Dotan Di Castro, and Shie Mannor.
Learning the variance of the reward-to-go.
Journal of Machine Learning Research, 17(13):1--36, 2016.

[4] Adam White, Joseph Modayil, and Richard S Sutton.
Surprise and curiosity for big data robotics.
In *AAAI-14 Workshop on Sequential Decision-Making with Big Data, Quebec City, Quebec, Canada*, 2014.

[5] Martha White and Adam White.
A greedy approach to adapting the trace parameter for temporal difference learning.
In *International Conference on Autonomous Agents & Multiagent Systems (AAMAS)*, pages 557--565, 2016.